Chapter 1

Frederick Boehm

February 4, 2019

Todo list

List of Figures

List of Tables

1. Complex traits & QTL mapping

- look at Karl & Saunak's chapter 1
- goal is to motivate study of complex traits with QTL mapping

Identification of genes that affect measurable phenotypes has a long and successful history in model organisms. Complex traits include classical clinical phenotypes such as systolic blood pressure and body weight as well as newly measurable biomolecular phenotypes like gene expression levels, protein concentrations, and lipid levels. Understanding the genetic underpinnings of such traits may inform many areas of biology, medicine, and public health.

The first reported QTL study is from 1923, 30 years before the discovery of the structure of DNA (Watson, Crick, et al., 1953). Sax (1923) examined seed weight for the common bean (*Phaseolus vulgaris*) in an F_2 intercross. He assigned each F_2 subject to a gene class by examining its seed color patterns.

Lander and Botstein (1989) kickstarted modern QTL mapping methods research with their seminal report in the late 1980s. Their article

Goals of a QTL study depend on its scientific context. Often a researcher seeks to identify genomewide positions of QTL for each trait of interest. In some studies, the total number of QTL for a trait may be more interesting than the QTL positions.

A QTL study begins with a scientific question and the choice of a study design. Elements to consider include the mating design, phenotyping plan, genotyping plan, and statistical analysis methods. For most of the last century, attaining clinical phenotypes, such as body weight, was much less costly than genotyping. This setting sometimes led researchers to selectively genotype only those organisms with extreme phenotype values. With diminishing costs for both genotyping and phenotyping, many recent studies genotype and phenotype all subjects.

Since the 1980s, researchers have written and shared computer software for QTL studies. Early efforts included MAPMAKER, QTL Cartographer,

Since the early 2000s, the "qtl" R package has been a state-of-the-art resource for QTL mapping studies. It is open-source, free to download, well documented, and well supported.

Mamm Genome. 1999 Apr;10(4):327-34. Overview of QTL mapping software and introduction to map manager QT. Manly KF1, Olson JM.

Mating designs for two-parent crosses

One widely used mating design is the backcross. Although variations are possible, one typically begins with two inbred lines. Let's designate the two lines as "A" and "B". Mating of lines A and B leads to offspring, which we designate as F_1 (to denote the first filial generation). The F_1 offspring, then, mate with the A line (ie, the parental line) to produce N_2 subjects, where the letter "N" denotes the offspring from a backcross and subscript 2 reflects the generation number.

We assume that we're working with diploid organisms, so that every subject has two copies - which need not be identical - of each chromosome. Let's assume, further, that we're working with mice, so that every organism has 20 chromosome pairs.

Inbred lines, by definition, are homozygous at all markers. Let's designate, for a given marker, the A line to have two copies of allele A and the B line to have two copies of allele B.

Let's consider the genetic makeup of the F_1 subjects. We first must examine gametogenesis, the process of producing gametes, or reproductive cells, in the parents. Gametogenesis results in production of haploid cells, *i.e.*, cells with only one copy of every chromosome. Two gametes, one from each parent, unite to form a diploid zygote. All F_1 subjects are genetically identical. For every chromosome pair, they inherited one copy of the A chromosome (from the A parent) and one copy of the B chromosome (from the B parent). In other words, every F$_1$ subject has genotype AB at every marker, where A the allele from the A parent and B the allele from the B parent. The N$_2$ generation, then, has subjects with either AA or AB genotypes at a given marker.

To understand whether a given subject has AA or AB genotype at a given marker, we need to consider the fact that gametogenesis involves crossover events before the diploid cells divide into haploid cells. A crossover event results in a swapping of DNA between the two copies of a chromosome. In inbred lines, this swapping of DNA between the two copies of a chromosome is undetectable with marker genotyping, because both chromosomes have the same allele (either two As or two Bs in our example). However, the F$_1$ subjects, with AB marker genotypes, have distinct alleles at every marker. Thus, marker genotyping has the potential to detect crossover events that occur during gametogenesis in the F$_1$ by examining marker genotypes in the N$_2$ subjects. A picture helps to clarify this idea (Figure **??**).

Add figure here AND add text explaining the figure here

Another widely used two-parent cross is the "intercross". In it, two inbred lines again mate to produce F$_1$ subjects. Then, however, two F$_1$ subjects mate to produce F$_2$ subjects. The crossover events that occur during gametogenesis in the F$_1$ subjects gives rise to the three marker genotypes observed in F$_2$ subjects: AA, AB, and BB. We present a diagram for the intercross in Figure **??**.

Quantitative traits

Quantitative traits typically are those that take continuous values over a (possibly infinite) interval. Distinct analysis methods - typically with statistical tools called "generalized linear models" - are needed for traits that take binary values or one of only a few discrete values or only whole numbers. Examples of quantitative traits include both clinical traits like body weight, height, systolic blood pressure, and fasting blood glucose level and newly measurable biomolecular traits like gene expression levels, protein concentrations, and lipid levels.

Statistical challenges in QTL studies

Broman and Sen (2009) articulate two statistical challenges: 1. the missing data problem and 2. the model selection problem.

The missing data problem arises because subjects in QTL experiments typically are genotyped at a set of discrete markers across the genome. They are not genotyped at every nucleotide base. The "missing" genotypes belong to those bases that are not genotyped. Statistical tools for probabilistically handling missing data, often with hidden Markov models,

what did Lander & Botstein do for missing data problem? How do they infer genotypes between markers?

QTL mapping in a backcross

QTL mapping in an intercross

Lander & Botstein 1989 Haley & Knott 1992 Martinez & Curnow 1992

Soller et al 1976

3. Multivariate QTL scan in two-parent crosses

Jiang & Zeng 1995 Knott & Haley 2000

Jiang and Zeng developed multivariate methods for QTL mapping in two-parent crosses. They devised a multivariate analog of Zeng's composite interval mapping (Zeng 1993, 1994). This strategy treats phenotypes as arising from a mixture of normal distributions in which each genotype class has distinct distribution parameters. Within this multivariate mapping framework, Jiang and Zeng developed the first test of pleiotropy vs. separate QTL.

Jiang and Zeng first developed a test for pleiotropy vs. separate QTL in two-parent crosses. They developed it in the context of their work in multivariate QTL studies with composite interval mapping. They framed the scientific question of whether two traits are affected by a single, shared locus or by two distinct loci in terms of two statistical hypotheses. The null hypothesis states that there is a single pleiotropic locus that affects both traits, while the alternative hypothesis is that there are two distinct but nearby loci, with each locus affecting exactly one trait.

Knott and Haley subsequently reported methods for testing pleiotropy vs. separate QTL in two-parent crosses. Knott and Haley integrated their earlier work on univariate QTL mapping with marker regression with Jiang and Zeng's multivariate methods to develop a test of pleiotropy vs. separate QTL with multivariate marker regression methods.

Jiang and Zeng (1995)

One disadvantage of multivariate compositive interval mapping is the computing requirements for the iterative expectation-maximization procedures needed for parameter estimation. This prompted Knott and Haley (2000) to develop a marker regression-based approximation to multivariate composite interval mapping. Knott and Haley (2000) used a multivariate linear model for simultaneous mapping of multiple traits. They also presented a multivariate marker regression-based test of pleiotropy vs. separate QTL. This test is suitable for subjects that are equally related to each other, like the collection of offspring in an F_2 intercross of two inbred lines.

4. Multiparental populations

- what are they? Breeding design for CC & DO. Why use them?

While QTL mapping studies in the 1990s contributed to many advances in genetics and biology, complex trait researchers recognized the mapping resolution limitations in crosses involving two inbred lines. Seeking greater precision for QTL positions, scientific communities collectively decided to pool their expertise and resources into community-supported and community-maintained model organism mapping populations. These new mapping populations would incorporate genetic material from more than two inbred founder lines. The accumulated meiotic recombination events over many generations would enhance mapping resolution over previously available populations.

Products of these community-based efforts include the Collaborative Cross and Diversity Outbred populations from mouse researchers and Drosophila Synthetic Population Resource (King et al., 2012) and in the fruit fly scientific community. In subsequent years, scientists created multiparental populations in many other organisms, including tomato, rice, maize (Lehermeier et al., 2014), wheat (Mackay et al., 2014; Huang et al., 2012; Milner et al., 2016), Arabidopsis (), apple (Allard et al., 2016)

5. Univariate QTL mapping in MPP

- contrast with univariate QTL mapping in two-parent crosses

6. Multivariate mapping in mpp

6A.

7. Testing pleiotropy vs separate QTL in MPP

Testing pleiotropy vs. separate

7A. allele effects plots to discern pleiotropy v separate QTL

King et al 2012 Macdonald & Long 2007 *maybe do a citation search on these 2 articles to see who has used their ideas* CAPE software package - what exactly is the CAPE method???

8. Testing pleiotropy vs separate QTL to dissect an expression trait hotspot Tian et al. 2016. ?Schadt et al. 2005?